PE0146

NCBI BLAST: Enhanced web usability through new result page, and effective genomic data access

Tao Tao, Peter Cooper, Scott Mcginnis, Wayne Matten, David Arndt, Greg Boratyn, and Tom Madden
Information Engineering Section, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)
8600 Rockville Pike, Bethesda, Maryland 20814, USA

Abstract

NCBI provides BLAST services through the BLAST homepage, where the public can search against the nucleotide database (NT), the non-redundant protein database (NR), a set of NCBI Reference Sequence Project (RefSeq) genomic sequences, and their annotated RNA or protein sequences. In this presentation, we will describe new features contained in the recently updated BLAST result pages that significantly enhance the page's usability. We describe the available RefSeq genomic databases and their relationship with genome records in NCBI's assembly database, which will enable more rational selection of databases and their taxonomic subset during a BLAST search. We will also cover alternative ways to access genomic sequences of interest, through assembly records on the web or FTP, for use in local standalone BLAST.

NCBI Web BLAST Services

- Reside at its own domain (blast.ncbi.nlm.nih.gov, 2006) with an reorganized homepage (2016, right)
- Offer easy access to core BLAST programs and a standard set of nucleotide and protein sequence collections as target databases
- Allow focused genomic data access through the BLAST Genomes search box
- Also provide access to specialized sequence analysis tools in "Specialized searches" section
- A well utilized set of services that serves over 50K users and does 350K interactive searches daily on average.

Search Forms & Key Functions

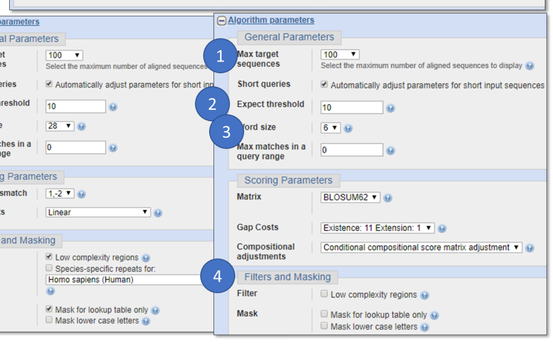
Databases and subset:

- Select the proper database from pulldown menu
- Specify a subset using Organism input & other preset filters (Entrez query no longer available for blastp)

Search stringency settings:

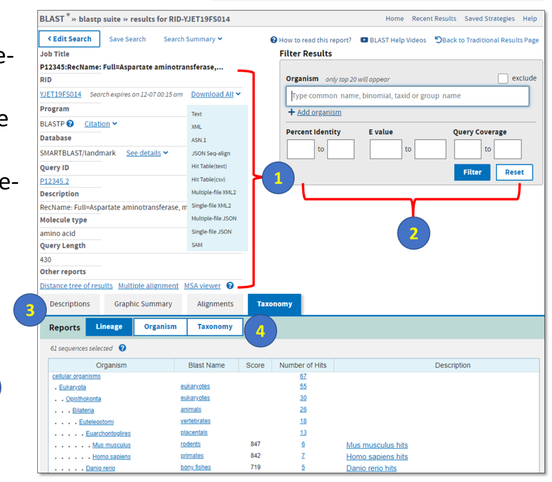
1. Upper limit of results via **Max target sequences**
2. Significance setting via **Expect**
3. Sensitivity via **Word size**
4. Suppression of spurious hits through **Filter** and **Repeat** masking (nucleotide only)

Last one is often overlooked by our users!

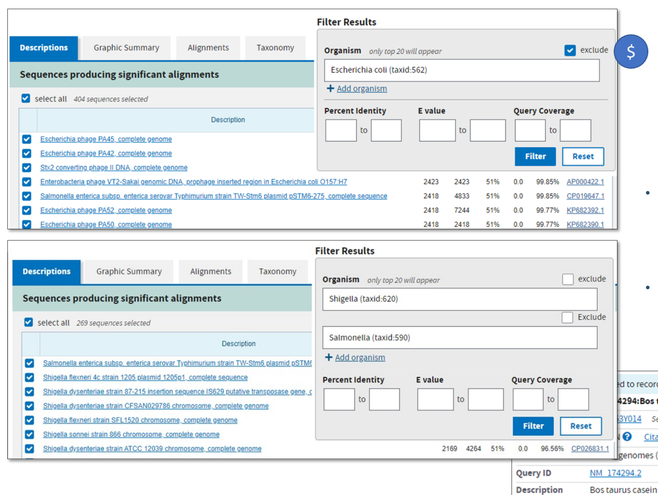


Updated Results Pages

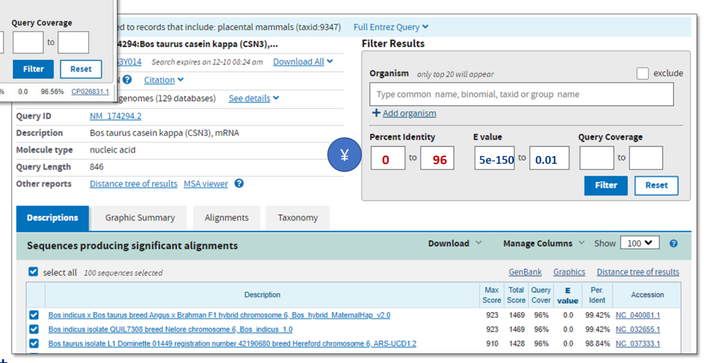
1. Places the **Search Summary** section at the top for convenient reference
2. Brings out the **result filtering functions** from hiding for ready access
3. Breaks results into **tabs** along the nature demarcation to reduce page scrolling
4. **Integrates Taxonomy report** as one of the tab sections [see right]
5. Makes other **reformatting functions context-sensitive** to reduce confusion (more on this later)



Filtering: an *E.coli* shiga toxin example



- Filter by exclusion (activated by "**exclude**" checkbox, $) to remove hits from the source organism that dominates the matches
- Filter by target organisms to see hits from interested sources. Multiple fields (added by "**+**" sign) operate in Boolean **OR** mode.

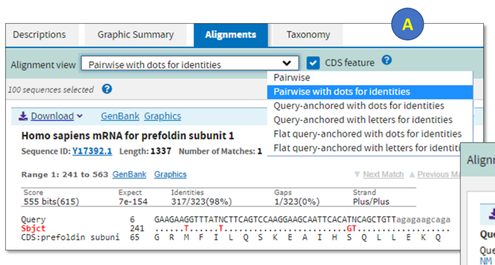


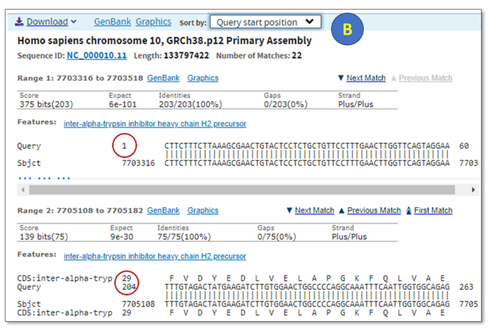
- Filter by **value ranges** (¥) to see desired subset of results satisfying the input criteria.
- Multiple ranges work in Boolean **AND** mode.

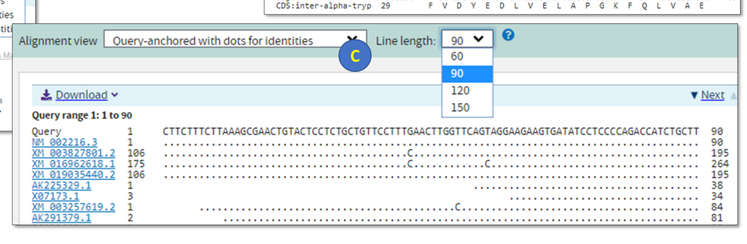
Responsive Design

Functions appear where they matter & relevant

A. Display format and CDS translation only in Alignment tab
B. Sort HSPs only appears for subjects with multiple matches to a query
C. Alignment length is only for pileup-like displays



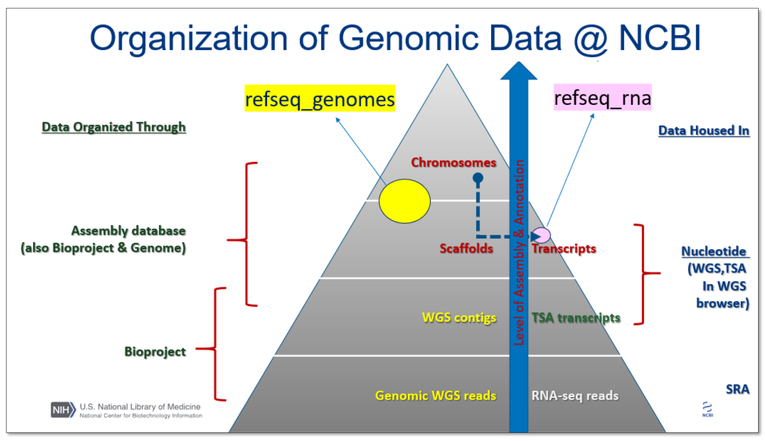




Nucleotide Databases

The figure to the right summarizes ways data (both sequence and metadata) are organized and stored at NCBI. Understanding it will help we locate needed data more effectively.

These two refseq databases are in the standard set of BLAST databases accessible from search pages listed in the Web BLAST section of the BLAST homepage.



Standard Nucleotide BLAST Databases

General purpose & default

- **NT:** Works fine for general purpose searches, but most eukaryotic genomes are NOT in it

Transcripts

- **Refseq_rna**: Curated or annotated mRNA and rRNA from NCBI RefSeq
- **TSA**: transcriptome assembly from submitters
- **SRA**: RNA-seq

Genomic Sequences

- **Refseq_genomic:** all genomic sequences (contigs, scaffolds, and full chromosomes), highly redundant
- **Refseq_genomes:** top-level genomic sequences, much less redundant
- **Refseq_representative_genomes:** selective genomes from the above, mostly one set for a given organism (prokaryotes are exceptions)
- **WGS:** whole genome shotgun contigs from submitters
- **SRA:** Raw nextgen reads

Genomic Data Available

Featured (6)

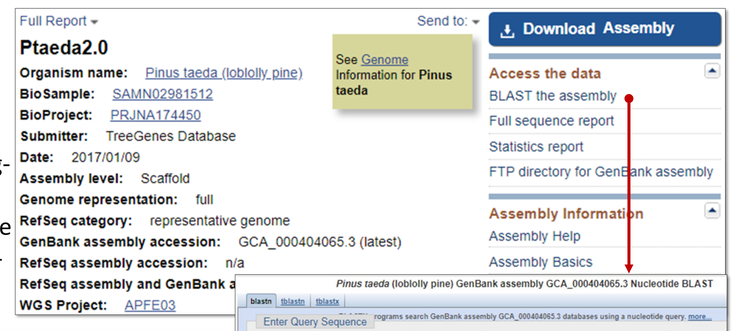
FTP - FTP Download B - Organism-specific BLAST AR - Annotation Report GDV - Genome Data Viewer

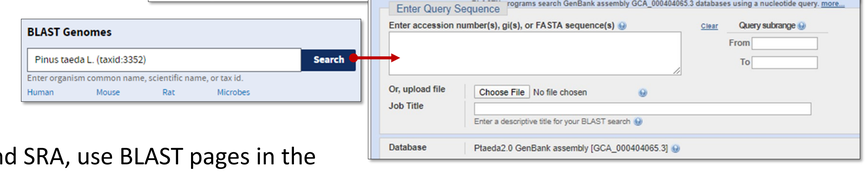
Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links
Homo sapiens (human)	GRCh38.p13 (GCF_000001405.39)	109.20191205	2019-12-05	2019-12-09	FTP B AR GDV
Mus musculus (house mouse)	GRCm38.p6 (GCF_000001635.26)	108	2019-07-12	2019-08-08	FTP B AR GDV
Rattus norvegicus (Norway rat)	Rnor_6.0 (GCF_000001895.5) Rn_Celera (GCF_000002265.2)	106	2016-07-07	2016-07-27	FTP B AR GDV
Danio rerio (zebrafish)	GRCz11 (GCF_000002035.6)	106	2017-06-02	2017-06-26	FTP B AR GDV
Apis mellifera (honey bee)	Amel_HAv3.1 (GCF_003254395.2)	104	2018-09-13	2018-09-19	FTP B AR GDV
Zea mays (maize)	B73 RefGen_v4 (GCF_000005005.2)	102	2017-12-07	2017-12-18	FTP B AR GDV

www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

The most informative way to get the summary and access the data is through the NCBI Eukaryotic Genome Annotation Pipeline's release list (partially shown above). The "**B**" link points to an organism-specific BLAST page, which provides access to the annotation release datasets, along with other related data for the organism.

For GenBank assemblies without refseq genome coverage , such as https://go.usa.gov/xpASP, their assembly records provide BLAST access to the original submitted assemblies. An easier way is to arrive at those organism specific genome BLAST page using the search box in the "BLAST Genomes" section of the BLAST homepage.





Access Other *less-Assembled* Data

To access datasets from WGS, TSA, and SRA, use BLAST pages in the "Web BLAST" section, set WGS, TSA, or SRA as the database, then apply organism or project-based limit. Using "WGS Project" or "TSA Project" option and the letter initial of the project offers more reliable performance. For SRA, experiment or run access is the only option. Searches against experiment with many large runs (SRR) could be difficult to run/complete. Local data access through SRAtoolkit is recommended.

Commandline Access to Genomic Data

This approach uses the **-remote** option of the standalone blast+ package. We need to know how to call the database of interest, and we can look up the database information through Entrez Programming Utilities or EntrezDirect:

```
$ esearch -db blastdbinfo -query 'ptaeda2.0' |esummary | xtract -pattern DocumentSummary -element Path
genomic/3352/GCA_000404065.3
```

The whole string (value in the <Path> field) is the input argument for the -db option:

```
$ blastn -remote -db genomic/3352/GCA_000404065.3 <...>
```

BLAST Access to Genomic Data Locally

Complete refseq_genomic set (large redundant) is available for download from the BLAST ftp site. Use the update_blastdb.pl to do this. Downloaded files are ready to use after extraction. Refseq_genomes download is planned for the near future.

For organism-specific subset, we can refer to assembly record for ftp path to the _genomic.fna.gz file for that assembly. We Need to format the sequence file into a blastable database using makeblastdb, before use. Batch downloading are described in the Assembly factsheet: ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_Assembly.pdf

Local Data Access Using SRAtoolkit

Searching against WGS, TSA, and SRA datasets may be necessary for organisms without fully assembled genome, scaffold, or good transcript collections. NCBI's SRAtoolkit provides a convenient venue to search against these datasets under a local setup or in a rented cloud instance:

- prefetch the datasets of interests from NCBI
- blast search them locally using blastn_vdb or tblastn_vdb

A factsheets provides more details and dockerized packages are also available:
ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_Local_SRA_BLAST.pdf, github.com/ncbi/docker

Visit NCBI Booth **321**

For our presentation slides write to us **info@ncbi.nlm.nih.gov**